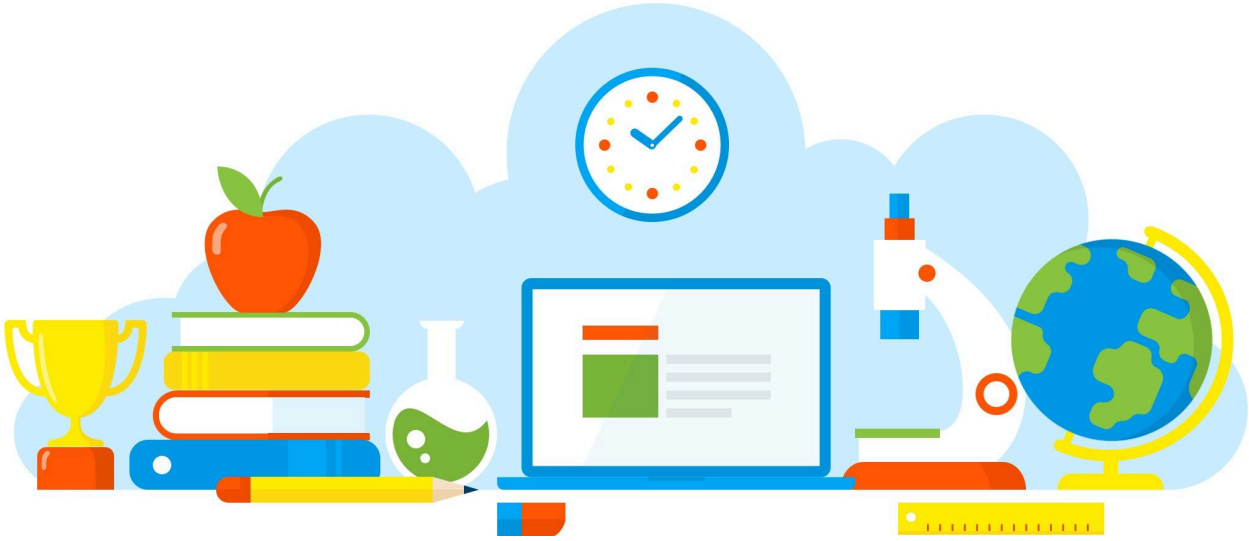

Bilan S1.04

Eddy Francou et Maxime Rastelli

Début et réflexions sur la base de données

La synthèse sur le naufrage du Titanic a permis de mettre en avant l'intérêt d'une base de données remplie avec les informations relatives au naufrage. En effet, cela permet de classifier les données et leur nature, et d'y accéder plus facilement grâce aux requêtes SQL. Cela permet aussi un certain ordre : les données ne sont pas triées en catégories, mais peuvent l'être selon le point de vue utilisé (la forme de la requête). Elle est ainsi nécessaire pour répondre à l'entièreté des questions que l'on se pose (comme une explication de la différence de taux de survie entre hommes, femmes et enfants).

Création de la base de données

Le SEA a été créé, puis transformé en SLR. Cela a permis d'identifier les facteurs importants de la base de données, de les mettre en avant, et d'en simplifier la structure. Ainsi il est moins nécessaire de réaliser des requêtes sur différentes relations en même temps. Les fichiers SQL de définition des tables et de peuplement permettent ainsi d'étudier les données du Titanic.

Comparaison des survivants et des victimes

Trois requêtes permettent de connaître le nombre de victimes et de survivants en fonction de leur classe. On remarque de manière générale que dans les classes inférieures, il y a eu une proportion de survivants par rapport aux victimes bien moins importante :

```
 survivants1 | victimes1
-------------+-----------
         200 |       123
(1 row)

 survivants2 | victimes2
-------------+-----------
         119 |       158
(1 row)

 survivants3 | victimes3
-------------+-----------
         181 |       528
(1 row)
```

Ce n'est d'ailleurs qu'en 1ère classe qu'il y a eu moins de victimes que de survivants.

On peut connaître aussi le nombre de survivants et de victimes par catégorie (enfant, femme, ou homme) :

```
 survivants_hommes | victimes_hommes
-------------------+-----------------
               109 |             501
(1 row)

 survivants_femmes | victimes_femmes
-------------------+-----------------
               267 |              79
(1 row)

 survivants_enfants | victimes_enfants
--------------------+------------------
                 51 |               39
(1 row)
```

Cependant ces données sont incorrectes, car dans l'instance actuelle de la base de données certains âges ne sont pas renseignés. Cela provoque un écart quand on compare la somme des survivants hommes femmes et enfants avec les survivants totaux, et pareil pour les victimes :

```
 survivants_somme | survivants_total
------------------+-----------------
              427 |             500
(1 row)
```

```
 victimes_somme | victimes_total
----------------+----------------
            619 |            809
(1 row)
```

Il convient alors de remplacer le jeu de données par une version dans laquelle tous les âges sont enregistrés.

De la même manière, on peut calculer automatiquement le taux de survivants.

Ensuite, on s'intéresse aux domestiques s'ils ont survécu et si oui, leur maîtres ont-ils aussi survécu?

Voici la liste des domestiques vivants et de leurs employeurs:

id_domestique	id_employeur	domestiques	employeur
709	306	1	1
346	499	1	0
1033	499	1	0
381	701	1	1
717	701	1	1
642	370	1	1
219	940	1	1
738	680	1	1
259	1235	1	1
843	764	1	1
291	988	1	1
538	1131	1	1
310	557	1	1
610	888	1	1
682	646	1	1
521	821	1	1
1306	308	1	1
1216	780	1	1
505	760	1	1
1267	956	1	1
951	1034	1	0
338	1088	1	1
1263	320	1	1
196	32	1	1
62	830	1	1
1048	1006	1	0
307	582	1	1
270	1206	1	1
966	1110	1	1

(29 rows)

On s’intéresse également au taux de survie par âge des enfants:

age	taux_de_survie
4	69
0	66
10	0
9	40
7	42
6	50
3	62
1	78
5	66
2	33
11	16
8	50

(12 rows)

On estime avec une requête que le nombre de survivants aurait pu être de 1178 au lieu de 500, ce qui signifie qu'on aurait pu sauver 678 personnes de plus:

```
 possible_survivants
---------------------
                 678
(1 row)
```